DOCUMENT RESUME

ED 281 854 TM 870 232

AUTHOR Raymond, Mark R.
TITLE An Interactive Approach to Analyzing Incomplete Multivariate Data.
PUB DATE Apr 87
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; *Data; Higher Education; *Multivariate Analysis; Predictive Measurement; *Predictor Variables; *Regression (Statistics); Research Problems
IDENTIFIERS Data Sets; *Missing Data

ABSTRACT

This paper examines some of the problems that arise when conducting multivariate analyses with incomplete data. The literature on the effectiveness of several missing data procedures (MDP) is summarized. The most widely used MDPs are: (1) listwise deletion; (2) pairwise deletion; (3) variable mean; (4) correlational methods. No MDP should be used without considering the characteristics of the data at hand, even though the literature may support one MDP. An application of these MDPs was demonstrated using college placement data taken from a published article on predicting success in foreign language training. The data included 12 variables (10 predictors and 2 criteria) for 279 college students. Descriptive statistics were produced by listwise and pairwise deletions, and missing predictor values were imputed. Missing data indicator variables were created to indicate the presence or absence of values on each of the original variables. The matrix of correlations between the indicator variables and the original variables appears in Table 3. The primary analysis involved the use of stepwise regression equations based on three alternative MDPs. The conclusion is made that regardless of the MDP selected, it is important to first determine the properties of the missing values, and second to decide which MDP is most appropriate. (BAE)

An Interactive Approach to Analyzing Incomplete Multivariate Data

Mark R. Raymond

American Nurses' Association

Paper presented at the annual convention of the American Educational Research Association, April 20-24, 1987, Washington, DC

AN INTERACTIVE APPROACH TO ANALYZING INCOMPLETE MULTIVARIATE DATA

Missing data can be a nuisance in all types of research, particularly in studies conducted in applied settings. The problem is compounded in multivariate research because nearly all forms of multivariate analysis require complete data from all cases; those cases with incomplete data are, by default, discarded. A casual review of any literature in the applied behavioral sciences, or of the most widely used statistical software packages, seems to indicate that the most widely used procedures for dealing with missing data are: listwise deletion, pairwise deletion, and mean substitution. Alternative procedures, such as imputing missing values by regressing variables with missing values onto relevant covariates, are rarely applied.

The purpose of this paper is to examine some of the problems that arise when conducting multivariate analyses with incomplete data. First, the literature on the effectiveness of several missing data procedures (MDPs) is summarized. Second, a rational application of these empirical MDPs is demontrated using college placement data. Specifically, it is suggested that no procedure should be blindly applied in any instance, even though the literature may strongly support the use of one MDP. Third, the results of regression analyses utilizing three different MDPs are compared using the same set of college entrance data. The motivation for this comparison is to determine the extent of the practical differences among MDPs.

The Effectiveness of MDPs

Numerous strategies for dealing with incomplete multivariate data have been proposed over the last 30 years. The most general and accessable of these MDPs are listed below.

1. LISTWISE DELETION. By default most multivariate algorithms discard all cases with incomplete data. The major limitation to this approach is that relevant data are frequently discarded. At best, power suffers and Type II error rates increase. At worst, bias is introduced.

2. PAIRWISE DELETION. This procedure utilizes all avaliable data in the computation of means and variances, and all avaliable pairs of values for the computation of covariances. Subsequent analyses (multiple regression) utilize the resulting summary statistics. A conceptual limitation is that when correlations and other statistics are based on different but overlapping subsamples of a larger sample, the population to which generalization is sought is no longer clear. Statistical problems may also arise when the data are missing in a nonrandom fashion. That is, it is possible to obtain correlation matrices with mutually inconsistent correlations.

3. VARIABLE MEAN. Missing values are replaced by the mean as computed for all cases that are present. The conceptual limitation with the variable mean MDP is that it is not intuitively appealing, or empirically accurate, to insert means when the missing observation in question occurs for a subject whose observations on other variables are quite distant from the mean. The empirical limitation is that inserting means will attenuate variance and covariance estimates.

4. CORRELATIONAL METHODS. Missing values are replaced with the an estimate that is obtained by regressing the incomplete variable onto one or more covariates. These procedures can be executed different ways by varying several factors: A) the starting point for computations (e.g., listwise deletion, pairwise deletion, etc.); B) the number of covariates used in the regression equations; C) whether or not an iterative solution is used; D) whether or not the resulting variance-covariance matrix is corrected for bias.

Several simulation studies have shown that obtaining estimates based on correlational procedures provides results that most closely approximate what would have been obtained had the data remained complete. The iterative regression algorithms have sometimes demonstrated quite dramatic improvements over listwise and pairwise deletion (e.g., Beale & Little, 1975; Gleason & Staelin, 1975; Chan, Gilman & Dunn, 1976). When the data are reasonably intact (better than 95% complete), and there are fewer than four variables, then the simple (bivariate) regression MDP is about as effective as the iterated regression procedure (Raymond & Roberts, 1987). A comprehensive review of the literature on multivariate analysis from incomplete data indicates that the most popular and expedient MDPs are, in general, the least effective (Raymond, 1986). In particular, pairwise deletion has been shown to lead to some very inaccurate results (Haitovsky, 1968).

An Approach to Analyzing Incomplete Data

Although there is ample support for imputing missing values through the use of correlational methods, these MDPs should be judiciously applied. No MDP should be used without first considering the characteristics of the data at hand. For example, using predictors to estimate criteria can result in inflated R^2s in subsequent analyses. Similarly, using two predictors to estimate one another may induce unwanted multicollinearity. Also, both the amount and the pattern of missing data may have implications for how subjects with missing values are treated. It may be necessary to drop some subjects while estimating missing values for other subjects. Similarly, if a particular variable is prone to extensive nonresponse, then that variable may need to be excluded from analysis.

The example below is intended to demonstrate some of the factors to consider when analyzing incomplete multivariate data. It is not intended to represent the correct approach, but rather one approach that recognizes the fact that missing data has the potential for altering the

conclusions of a study. The remainder of the paper is presented in three sections: the description of the data set, the description and treatment of nonresponse, and the execution of the primary analysis. The primary analysyis involves a comparison of stepwise regression equations computed from data treated by three MDPs.

Data set

The data originally appeared in a published article on predicting success in foreign language training. The raw data included 12 variables (10 predictors and 2 criteria) for 279 college students. The primary purpose of the analysis was to determine if the prediction of performance in a foreign language course could be augmented by adding one or more new variables to the existing prediction system. While not all 12 variables are directly pertinent to this analysis, all will be useful in the treatment of the missing values.

The variables include: high school GPA (HS-GPA), SAT-V and SAT-M scores, English placement test scores, two subtest scores on the Modern Language Aptitude Test (MLAT-4 and MLAT-5), Foreign Language Attitude scores, age, sex, prior experience with a foriegn language, current college GPA, and final grade in a foreign language course (FL Grade).

Description and explanation of nonresponse

The total sample size included 279 cases. Due to missing data on one or more of the variables, the listwise deletion sample size consisted of 174 subjects, or 62% of the original sample. The pairwise deletion sample sizes ranged from 197 to 278. Most of the missing data occurred for MLAT-4, MLAT-5, and FL Grade. In all about 9.2% of the original data matrix was missing (N=279, p=12, m=308).

Of the 279 subjects, two of the students were missing data on six of the ten predictors. These two students were elimated from any further analysis. Another 45 students had incomplete criterion data (i.e., no data for FL Grade), and were excluded. Most of these students were audits or drop-outs. Thus the sample size was reduced to N=230. This revised data set was 93.9% complete (based on the revised sample of N=230, p=12), as it still consisted of 168 missing values. This revised data set was saved and will be restored to completeness by estimating the missing values. It is important to note that the entire data matrix (N=279) will not be restored; only the subset (N=230) with complete data for FL Grade.

Missing values on FL Grade were not to be estimated from certain predictors because the R^2 from the primary analysis would be inflated. That is, using MLAT scores to estimate missing values on FL Grade would only serve to artificially increase the relationship between the two variables in the final analysis. Further, it did not seem appropriate to estimate FL Grade from College GPA (an irrelevant criterion variable), since the relationship of FL Grade with certain predictors may have been lowered, specifically those predictors designed to be orthogonal to college GPA while being related to FL Grades. It is worth noting that students with missing FL Grades performed slightly lower

than other students on HS-GPA, MLAT-4, and SAT-M. The likely consequence of eliminating these students will be a restriction in range, which may attenuate some of the correlations.

A portion of Table 1 lists the descriptive statistics produced by listwise and pairwise deletion. A comparison of correlation matrices reveals some differences in the maginitude of the correlations. For example, the correlation between MLAT-4 and FL Grade was .38 using listwise deletion and .43 using pairwise deletion. There were also some differences in means and standard deviations produced by the two methods. Pairwise deletion resulted in lower means and smaller standard deviations produced by on some of the ability predictors suggested that lower ability students were more likely to have missing data. Listwise deletion seemed to be an undesirable alternative for two reason: loss of power and potential bias. Pairwise deletion did not seem to be the desirable solution for theoretical reasons related to sampling and generalization and because of its poor performance in prior simulation experiments.

At this point in the analysis, the imputation of missing predictor values seemed a reasonable course of action. Initial estimates were first obtained using college-GPA, an otherwise irrelevant variable. College GPA was moderately correlated with the ability predictors, and thus provided more accurate initial estimates than would have been provided by the inserting the variable mean. Next, the initial estimates were revised using stepwise regression for two iterations. In all, 168 blanks on eight variables were replaced by estimates, with the eight multiple Rs ranging from .27 to .79, with a mean multiple R of about .60. The descriptive statistics provided by the restored data matrix also appear in Table 1. Again, there are some notable differences among correlations produced by the three MDPs. A checkmark (✓) indicates correlations that differ by at least ± .05.

As an aide to the further investigation of nonresponse, missing data indicator variables were created to indicate the presence or absence of values on each of the original variables. For example, if a subject failed to respond to the item concerning Age, they were assigned a value of 1 on a new variable called #Age. The descriptive statistics for the missing data indicator variables revealed a few important characteristics of the data. The sum of observations (1 or 0 over N subjects) corresponded to the number of missing values, while the mean reflected the proportion of students with missing data. The correlations indicated whether or not there was a general tendency for nonresponse to some of the variables. As indicated in Table 2, some of the correlations among the indicator variables were .98 and 1.00.

> [AN ASIDE: A principal components analysis of the matrix in Table 2 resulted in a pattern of loadings that corresponded to the manner in which the data for this study were collected. The data for this study were collected from two sources: College records and measures collected during two sessions that occured during class time. The loadings on the first component corresponded to missing data on variables obtained from college records. Variables loading on the second

component were those collected during the two class sessions (MLAT scores, demographics and attitudes). Missing data on these variables occured for students who elected not to participate in the study. The third component was bipolar, and corresponded to students who provided demographic and attitudinal information but did not take the MLAT. The principal components analysis was not necessary in the present context because the data set was small enough, and the missing values few enough, to obtain the same information by a visual inspection of the missing data matrix. However, such an analysis would be quite helpful for large data sets.]

Table 3 presents the matrix of correlations between the indicator variables and the original variables. Several moderate correlations suggested that older students, and students scoring lower on many of the academic performance measures were more likely to have missing values. A three variable regression equation consisting of Age, MLAT-4, and College GPA predicted "missingness" on college records (SAT, HS-GPA and English Placement) with a multiple correlation of .47. It is obvious that the data are missing in a nonrandom fashion; thus generalizability of the regression equations obtained from the primary analysis will be limited.

Execution of Primary Analysis

The primary data analysis involved the use of stepwise regression to determine the effectiveness of ability, demographic, and attitudinal variables in predicting performance in foreign language courses. Here we will compare three regression equations based on three alternative MDPs: pairwise deletion (N=174), listwise deletion (278 <N< 197), and restoration by imputing missing values (N=230). The data based on listwise deletion resulted in a five variable equation that produced an R^2 of .35. Pairwise deletion produced a four variable equation with an R^2 of .29. The restored data gave a different arrangement of the same variables which produced an R^2 of .30. The standardized regression equations appear below.

LISTWISE DELETION: R^2=.354

FL Grade = HS-GPA(.351) + FL Att(.211) + Age(.252) + MLAT-4(.231) + Prior Exp(.164)

PAIRWISE DELETION: R^2=.291

FL Grade = MLAT-4(.275) + HS-GPA(.262) + FL Att(.192) + Prior Exp(.123)

RESTORATION: R^2=.299

FL Grade = HS-GPA(.309) + FL Att(.196) + MLAT-4(.252) + Age(.163) + Prior Exp(.147)

Concluding Comments

Although it may be argued that no analysis should be conducted when the level of missing data exceeds a certain percentage (say 5%), the fact remains that such studies are often reported in the literature. Conducting analyses on incomplete data can be justified by the argument that the less than precise information provided by studies utilizing fragmentary samples is preferable to no information at all.

The three MDPs provided somewhat different findings. The interpretation of the reults should be guided by the characteristics of the missing data. The correlations in Table 3 can be useful for this. Specifically, the results may have limited applicability to older students. The prediction model developed from listwise deletion completely ignored the students with incomplete data (many of whom were older or lower in academic aptitude). Therefore, any report based on the listwise deletion MDP should mention this limitation. It is difficult to specify how the results based on pairwise deletion should be interpreted. The level of incompleteness (9.2%), and the systematic nature of the missing data would urge an extremely cautious interpretation of the regression equations.

The generalizability of the regression equations based on the restored data matrix depends on the accuracy with which the missing values have been imputed. At a minimum, there will be a restriction of range due to the imputed values being regressed toward the mean. Corrections for this type of systematic error are discussed in Beale and Little (1975), and Little (1978). Even more serious errors may result if the relationships between the incomplete variables and those variables used to estimate the missing values are not linear over the full range of values. In such instances, missing values may be drastically underestimated or overestimated. Regression diagnostics can be useful for assessing the impact of the imputed values on the prediction models obtained from restored data matrices. In the current restored data, there was a tendency for those cases with missing values to be overpredicted by the regression model (a preponderance of negative residuals). In addition, there were three cases flagged with high leverage indexes. Each of these three cases were individuals with incomplete data.

Which MDP should be trusted for the current analysis? While the results of several Monte Carlo studies offer fairly strong support for the use of correlational methods to impute missing values, one may be inclined to feel a little uncomfortable with fabricating data, even if it is done empirically. Adopting such a position requires that some other meaningful strategy be used to reach a decision. (We can readily acknowledge that choosing the MDP that gives the highest value of R^2 is not acceptable!) Examining the regression statistics for the three MDPs for consistency may provide some indication of which MDP should be chosen, or which should be avoided.

The differences in the regression weights between listwise deletion and restored data were minor, even though the R^2 values were notably different. On the other hand, pairwise deletion and estimation by

regression provided similar values of R^2, even though regression equation provided by pairwise deletion was substantially different from the regression equations provided by the other two MDPs. Given that the goal in conducting regression analyses on college entrance data is to develop empirical models which can be applied to future samples, it seems that more credence should be given to the consistency of beta weights. This line of reasoning rules out pairwise deletion. Choosing between listwise deletion and the restored data should be a function of the population to which generalizability is sought, and the degree of bias that may have been introduced by each of the two MDPs.

Regardless of the MDP selected, it is important to first determine the properties of the missing values, and second, to overtly decide which MDP seems most prudent. Clearly, no MDP is a perfectly reliable substitute for complete data. However, when missing data do occur, it is usually the case that some MDP is utilized, even if by default. Listwise deletion and pairwise deletion, while sometimes appropriate, should not be relied on simply as a matter of expedience.

References

Beale, R.M.L. & Little, R.J.A. (1975). Missing values in multivariate analysis. Journal of the Royal Statistical Society, B, 37, 129-145.

Chan, L.S., Gilman, J.A. & Dunn, O.J. (1976). Alternative approaches to missing values in discriminant analysis. Journal of the American Statistical Association, 71, 842-844.

Gleason, T.C. & Staelin, R. (1975). A proposal for handling missing data. Psychometrika, 40, 229-252.

Haitovsky, Y. (1968). Missing data in regression analysis. Journal of the Royal Statistical Society, B, 30, 67-82.

Little, R.J.A. (1978). Consistent regression methods for discriminant analysis with incomplete data. Journal of the American Statistical Association, 73, 319-322.

Raymond, M.R. (1986). Missing data in evaluation research. Evaluation & the Health Professions, 9, 395-420.

Raymond, M.R. & Roberts, D.M. (1987). A comparison of methods for treating incomplete data in selection research. Educational & Psychological Measurement, 47, - .

Table 1

DESCRIPTIVE STATISTICS OBTAINED FROM THREE MISSING DATA STRATEGIES

	Age	Sex	Prior Exper	FL Atti	MLAT 4	MLAT 5	HS GPA	SAT-V	SAT-M	Eng Place	FL Grade	College GPA	N	MEAN	S D
Age	100												174	1.62	0.65
	100												268	1.72	0.87
	100												230	1.73	0.83
Sex	-25	100											174	0.46	0.50
	-24	100											278	0.45	0.50
	-23	100											230	0.45	0.50
Prior Exper	-06	13	100										174	3.19	1.60
	-17✓	14	100										268	3.21	1.60
	-15	14	100										230	3.23	1.58
FL Attitude	-15	30	06	100									174	83.16	12.64
	-13	27	02	100									279	82.49	14.03
	-12	29	02	100									230	83.21	13.09
MLAT - 4	-11	16	-11	11	100								174	25.31	6.11
	-17✓	17	-06✓	16✓	100								230	24.26	6.23
	-22	19	-08	13	100								230	24.67	5.92
MLAT - 5	-09	17	-09	12	22	100							174	19.69	4.85
	-20✓	17	-02✓	11	25	100							230	19.24	4.92
	-22	18	-07	11	25	100							230	19.37	4.63
HS GPA	-14	23	05	06	50	32	100						174	3.36	0.46
	-14	23	08	05	49✓	32	100						245	3.29	0.48
	-18	23	05	07	54	35	100						230	3.31	0.45
SAT - V	07	-01	-09	09	43	23	26	100					174	507.36	89.20
	03	-04✓	-06	04✓	42✓	17✓	27✓	100					245	501.51	91.16
	07	-07	-09	03	37	17	22	100					230	507.53	86.20
SAT - M	00	-22	-11	-15	41	12	30	35	100				174	575.93	85.32
	-02	-22	-06✓	-08✓	44	12	35✓	41✓	100				245	564.25	88.71
	-05	-23	-11	-12	42	13	32	36	100				230	570.71	83.61
Eng Place	-06	12	-08	19	58	25	32	74	40	100			174	54.92	15.60
	-08	11	-06	15	58	23	32	73	42	100			242	53.95	15.40
	-10	10	-08	18	59	23	32	72	39	100			230	54.39	14.89
FL Grade	14	15	15	23	38	21	45	18	19	25	100		174	3.40	0.85
	-02✓	18	13	25	43✓	20	42	10✓	15	18✓	100		232	3.32	0.88
	01	17	12	23	40	20	44	10	16	21	100		230	3.34	0.87
College GPA	16	02	-09	07	40	34	46	31	24	33	58	100	174	2.84	0.61
	05✓	03	-05✓	06	42	29✓	48	29✓	30✓	30✓	57	100	278	2.75	0.62
	11	02	-11	05	38	31	48	25	25	28	57	100	230	2.77	0.62

Listwise Deletion

Pairwise Deletion

Estimation by Linear Regression

Table 2

CORRELATIONS BETWEEN MISSING DATA INDICATOR VARIABLES

	#Age	#Sex	#Prior Exper	#FL Atti	#MLAT 4	#MLAT 5	#HS GPA	#SAT-V	#SAT-M	#Eng Place	#FL Grade	#College GPA	Percent Missing	Number Missing
#Age	100												3.5	8
#Sex	.	.											0.0	0
#Prior Exper	100	.	100										3.5	8
#FL Attitude	.	.	.	.									0.0	0
#MLAT - 4	14	.	14	.	100								12.6	29
#MLAT - 5	14	.	14	.	100	100							12.6	29
#HS GPA	10	.	10	.	-13	-13	100						10.0	23
#SAT - V	10	.	10	.	-13	-13	100	100					10.0	23
#SAT - M	10	.	10	.	-13	-13	100	100	100				10.0	23
#Eng Place	09	.	09	.	-13	-13	98	98	98	100			10.4	24
#FL Grade	.	.	.	.	.	.	.	.	.	.	.		0.0	0
#College GPA	-01	.	-01	.	-03	-03	20	20	20	19	.	100	0.4	1

Table 3

CORRELATIONS BETWEEN ORIGINAL VARIABLES AND MISSING DATA INDICATOR VARIABLES

	Age	Sex	Prior Exper	FL Atti	MLAT 4	MLAT 5	HS GPA	SAT-V	SAT-M	Eng Place	FL Grade	College GPA
#Age	-05	07	04	12	02	05	-03	-05	00	03	-05	-08
#Sex	.	.	.	.	.	.	.	.	.	.	.	.
#Prior Exper	-05	07	04	12	02	05	-03	-05	00	03	-05	-08
#FL Attitude	.	.	.	.	.	.	.	.	.	.	.	.
#MLAT - 4	-03	-00	09	-05	-04	-03	-11	04	-07	-00	00	-11
#MLAT - 5	-03	-00	09	-05	-04	-03	-11	04	-07	-00	00	-11
#HS GPA	41	-04	-02	-01	-26	-18	-16	-00	-09	-09	-23	-15
#SAT - V	41	-04	-02	-01	-26	-18	-16	-00	-09	-09	-23	-15
#SAT - M	41	-04	-02	-01	-26	-18	-16	-00	-09	-09	-23	-15
#Eng Place	39	-05	-04	00	-26	-19	-17	01	-07	-08	-22	-13
#FL Grade	.	.	.	.	.	.	.	.	.	.	.	.
#College GPA	02	-06	-09	06	02	-05	-02	01	02	01	-03	-01

NOTES: 1. Indicator variables are those preceded by a # symbol. 0 = DATA PRESENT, 1 = DATA MISSING.
2. Correlations in Tables 2 and 3 are based on N = 230 individuals with final grades in FL course.
3. The missing entries in Tables 2 and 3 are due to complete data for SEX, FL ATTITUDE, and FL GRADE for the 230 cases.
4. Prior to computing correlations in Table 3, missing values were replaced by estimates obtained from multiple regression. Without estimates it would not have been possible to obtain the correlations that occupy the diagonal (e.g., a correlation between age and #age).